SGLang: Efficient Execution of Structured Language Model Programs

Year: 2023 | *Citations: 401* | *Authors: Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie*

Abstract

Large language models (LLMs) are increasingly used for complex tasks that require multiple generation calls, advanced prompting techniques, control flow, and structured inputs/outputs. However, efficient systems are lacking for programming and executing these applications. We introduce SGLang, a system for efficient execution of complex language model programs. SGLang consists of a frontend language and a runtime. The frontend simplifies programming with primitives for generation and parallelism control. The runtime accelerates execution with novel optimizations like RadixAttention for KV cache reuse and compressed finite state machines for faster structured output decoding. Experiments show that SGLang achieves up to 6.4x higher throughput compared to state-of-the-art inference systems on various large language and multi-modal models on tasks including agent control, logical reasoning, few-shot learning benchmarks, JSON decoding, retrieval-augmented generation pipelines, and multi-turn chat. The code is publicly available at https://github.com/sgl-project/sglang